DOCUMENT RESUME

ED 076 669 TM 002 677

AUTHOR Bernstein, Ilene Nagel; And Others
TITLE Should Evaluation Researchers in Education Have an Inferiority Complex?
SPONS AGENCY Russell Sage Foundation, New York, N.Y.
PUB DATE 73
NOTE 29p.; Paper presented at annual meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Comparative Analysis; *Educational Research; Educational Researchers; *Evaluation; *Federal Programs; *National Surveys; *Organization; Research Methodology; Speeches

ABSTRACT

Evaluation has been received by social scientists and administrators alike as the critical step required for the identification and implementation of action programs that are both effective and within the resources of the country. Very little is known about the actual evaluation process, its management methodologies and its impact. A national survey was conducted at Russell Sage Foundation by the authors of all federally funded evaluation studies in the human resource area. Studies were included if they were initiated in fiscal 1970 and had a budgetary allotment of $10,000 or more. The findings of this survey are described. Comparisons are made with respect to both process and impact evaluation between investigations in education and those in other fields such as health, income security, public safety, welfare, etc. In addition to comparing the characteristics of education studies to those of others, data on the organizational arrangements under which the evaluation is carried out and the characteristics of the researchers are presented by field. (Author)

FORM PAGE 10
PRINTED IN U.S.A.

FILMED FROM BEST AVAILABLE COPY

ED 076669

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Should Evaluation Researchers In Education Have An Inferiority Complex?*

Ilene Nagel Bernstein
University of Minnesota

Howard E. Freeman
Ford Foundation

George W. Bohrnstedt
University of Minnesota

TM 002 677

*Paper presented at the American Educational Research Association Meetings, New Orleans, 1973.

Should Evaluation Researchers In Education Have An Inferiority Complex?*

Ilene Nagel Bernstein
University of Minnesota

Howard E. Freeman
Ford Foundation

George W. Bohrnstedt
University of Minnesota

Evaluation research has been defined as the application of the scientific method for assessing the effectiveness of an activity in producing some desired social goal.[1] While the definition varies from author to author, the essential emphasis on the utilization of sound empirical methodology as opposed to non-systematic testimonial-like data remains constant. And yet, despite the encouragement for at least a decade by leading scholars in all academic disciplines, no one as yet has stood up to praise the efforts of evaluators or to acclaim their studies as models of good research. With few exceptions, most evaluation research is still described as at best 'lacking' and at worst 'the major contributor to a Journal of Irreproducible Results.' The explanations range from a lack of adequate available methodological techniques through problems inherent in the nature of the research because of its politicality, to the poorly trained persons who do evaluation research. The relevant literature abounds with discussions of methodological problems and assertions about the conditions under which methodological adequacy is strengthened or weakened,

---

*This research was supported by a grant from the Russell Sage Foundation as part of a larger study on the Collation and Dissemination of Evaluation Research.

and yet to date there have been no systematic studies of this activity which engages so many professionals. Accordingly, it was our contention that two basic questions had to be addressed; the first was to ascertain something about the state of the art of evaluation research, i.e. where does the support come from, in what form, to whom are awards made, what kinds of organizations are awardees affiliated with, what is the academic background of the evaluation staff, and what are some of the structural conditions under which evaluations are carried out.[2] The second was to ask, given that the quality of evaluation research varies, what are the factors which might possibly account for some of that variance. While we could not make a definitive assessment of quality because of the limitations of our data, we could make some assertions based on rigorous analyses of the relative quality of research for a population fairly large in size. Additionally, by using the data we collected for the descriptive part of our studies, we could test certain assertions about the variation in quality as it relates to such dimensions as nature of award, sponsoring agency, organizational arrangements, academic discipline of staff, and the like.

The research reported herein represents a part of a more comprehensive study of all evaluation research funded in fiscal 1970, however we focus here on academic discipline as a major variable, specifically looking at evaluators in education. Educationalists as well as evaluations focusing on education are particularly interesting for several reasons. First, perhaps the most controversial evaluation in recent years or at least the one with greatest publicity, has been the evaluation of an educational program, i.e. Head Start. One of the many consequences of the 'Head Start' evaluation was to bring to the forefrontmethodological debates about evaluation research, as well as a flurry of general discussions about this heavily funded, politically relevant,

socially needed, but little known about activity in which so many persons are engaged. Second, education has long been thought of as the major means by which persons achieve social mobility. Since inequality was considered, at least by recent Democratic administrations, to be the single greatest problem to/overcome/to achieve the Great Society, education was a natural setting for reform experiments. Thus, it is an area which because of its social importance has special interest.

Summarily, the following presentation attempts to describe the state of the art of evaluation research for persons specializing in education, i.e. educationalists as well as assessing how they fare relative to persons from other academic disciplines in the quality of their research. In order to provide a broader basis for comparison, supplementary data are cited as well, especially with respect to factors relating to 'quality of research.'

Since ours was an exploratory study, our hypotheses would be better categorized as 'generally stated assertions', some of which were theoretically based, others more experiential or stemming from fundamental beliefs about the social reality of the field of evaluation research. In any case, we asserted that the evaluation studies would vary considerably in the quality of their research and that that variation would be related to the size of the budget, the nature of the award, the length of time for the evaluation, the federal agency sponsoring the research, the type of organization conducting the research, the conditions under which the research was carried out, and so on. As for the direction of these assertions, i.e. which category would do better, in almost all cases there were contradictory assertions in the literature. For example, some posited that since good research is costly, those with larger budgets would fare better. Others on the other hand assert that, besides the fact that much of the best social scientific research has been done with small budgets, large budgets are often allocated to evaluations of programs with loosely developed ideas in the hopes that somehow such a

huge sum of money will not only provide for an adequate evaluation but simultaneously will straighten out the ill-defined action program as well. On the more cynical side, some assert that large budgets are allocated by agencies to signify their sincerity, but in fact camouflages what may be in fact an attitude of disinterest in any evaluation results, regardless of findings. As such, they imply awards are given knowingly to incompetent researchers whose results can then be discredited because of methodological inadequacies. While we could continue to present the conflicting assertions for each of the variables named and make a prediction for each, we shall opt instead to explore the results rather than to posit hypotheses as such. Following the advice of Michael Scriven, one of education's leading contributors to the evaluation research literature, we will treat our "evaluation of evaluation research" as a formative evaluation and proceed accordingly.[3]

Before commencing with the presentation of our data, a brief description of the procedures used for sample selection and methods of data collection seems in order.[4] Our population of interest included all evaluation studies funded directly by the agencies of the federal government in the fiscal year 1970. Accordingly we obtained a list from each agency of every award given in FY 70 for an evaluation of a large-scale social action program, aimed at ameliorating some social problem, in the areas of health, education, welfare, income security, public safety, housing, and manpower, and, with a minimum research budget of $10,000 or more. All persons on that list were then sent and asked to complete a copy of our questionnaire. Eighty-four per cent of those persons returned a completed questionnaire (N = 318). Of those, 74% (N = 236) responded that they indeed had done an evaluation study and 26% (N = 82) responded they had not.[5] Thus the results presented below are based on data compiled from those 236 respondents.

At the outset, let us cite some very basic descriptive facts about: a) evaluations done by persons with degrees in education, b) evaluations

focusing on education as a topic and c) evaluations sponsored specifically by the federal agency most concerned with education, the Office of Education. Our data indicate that of all evaluation research funded in fiscal 1970, 11% was awarded to persons who indicated that their academic discipline or area of specialization was education. With respect to evaluations which focused on education as the primary concern, we found that 16% of all studies were so categorized. Lastly, when looking at the various agencies and their relative contribution, we find that the Office of Education sponsored 10% of all evaluation research in fiscal 70. The variation in these three figures reflects not a discrepancy in the data but rather that: a) not all persons specializing in education study problems focusing on education, b) evaluations focusing on education were done by persons within a variety of academic disciplines, i.e. 8.1% by economists, 29.7% by psychologists, 2.7% by sociologists, 16.2% by 'others', and 43.2% by those specializing in education, and c) the Office of Education awarded research funds to persons specializing in disciplines other than education as well, especially including psychologists. This brings us to an important point to note which is that 19% of the psychologists were working on evaluations focused on education. This coupled with the fact that educational psychologists often/probably categorized themselves as psychologists rather than educationalists leads us to suspect that a truer estimate of evaluation research being done by educationalists would be closer to 15% than the 11% figure cited previously. While this may seem trivial now, it will become important as a factor to recall later on in our discussions of academic discipline and its relationship to the quality of evaluation research. Unfortunately in the analyses which follow we will use only those 11% who stated their academic discipline was education since we have only inferential and not concrete evidence to the contrary.

Turning our attention first to the source of funding we find that educationalists received their awards for evaluation research primarily from Social Rehabilitation Services (SRS) (32%) and from the Office of Education and other HEW monies, (Not including NIH, NIMH, SRS) (36%). This was not an unexpected result. Perhaps of greater interest is that these awards tended to be larger rather than smaller, i.e. 64% of their awards were for $100,000 or more as compared to 49% for all researchers combined.

Since the amount of the financial stipend is closely related to the nature of the award, i.e. contracts tend to have larger budgets than grants, we examined the relationship between academic discipline and nature of award and found that educationalists tended to work slightly more on grants (56%) than on contracts (44%). Interestingly however, psychologists do vary substantially with 74% of their work being supported by grants as opposed to 26% supported by contract.

The last variable of interest in this section was the amount of time allotted for the evaluation. While we know from our data that this is related to both the size of award, i.e. longer studies carry larger financial stipends, and to the nature of the award, i.e. longer studies are more likely grants rather than contracts, we thought it useful to see if there was any systematic variation with respect to academic discipline, i.e. 'educationalists' tend to be represented in roughly equal proportions with respect to the amount of time, however for psychologists there was a very strong tendency to work on the longer studies with respect to temporal duration. This however is closely related to the fact that psychologists also tend to work more on grants and on studies having large financial stipends.

Moving on to the second area of interest let us describe something about the characteristics of the organizations with which the awardees were affiliated. First, with respect to the type of organization which was conducting the evaluation, we find again that educationalists do not vary as much as the other disciplines across the different organizational types. While the numbers

in the cells are small, it is still surprising to note that 28% of the educationalists are located in profit-making corporations. The 24% in public or service agencies are not as unlikely since they probably represent persons working in public departments of education. In order for you to see the distribution of educationalists versus other disciplines this relationship is shown in Table 1.[6]

Table 1

Type of Organization and Project Director's Academic Discipline

A. Percentaging by Type of Organization

Type of Organization	Educ.	Psych.	Econ.	Sociol.	Med., S.W. and Psycr.	Other	Totals
Profit	11.3(7)	6.5(4)	29.0(18)	14.5(9)	1.6(1)	37.1(23)	100.0(62)
Non-profit and Research	9.6(5)	25.0(13)	15.4(8)	9.6(5)	13.5(7)	26.9(14)	100.0(52)
Educational Institution	8.6(7)	34.6(28)	9.9(8)	12.3(10)	19.8(16)	14.8(12)	100.0(81)
Public/Service Agency	15.8(6)	31.6(12)	10.5(4)	10.5(4)	15.8(6)	15.8(6)	100.0(38)
Totals	10.7(25)	24.5(57)	16.3(38)	12.0(28)	12.9(30)	23.6(55)	100.0(233)

B. Percentaging by Discipline of Project Director

Degree of Project Director	Profit	Non-profit and Research	Educational Institution	Public/Service Agency	Totals
Education	28.0(7)	20.0(5)	28.0(7)	24.0(6)	100.0(25)
Psychology	7.0(4)	22.8(13)	49.1(28)	21.1(12)	100.0(57)
Economics	47.4(18)	21.1(8)	21.1(8)	10.5(4)	100.0(38)
Sociology	32.1(9)	17.9(5)	35.7(7)	14.3(4)	100.0(28)
Med., S.W., & Psycr.	3.3(1)	23.3(7)	53.3(16)	20.0(6)	100.0(30)
Other	41.8(23)	25.5(14)	21.8(16)	10.9(6)	100.0(55)
Totals	26.6(62)	22.3(52)	34.8(81)	16.3(38)	100.0(233)

Note -- $\chi^2_{15} = 41.4$, $p < .001$. These cases omitted because of blank responses.

In terms of an overall picture then, thus far we have noted that educationalists receive the greatest proportion of their funding from SRS and the Office of Education, the form of award being nearly evenly divided between grants and contracts with slightly more of the former, and with budgeting allotments larger in size than smaller. Further, they tend to be found almost equally in all types of organizations, and show no propensity to do studies longer or shorter in time.

Turning our attention to the conditions under which educationalists work, we will describe three specific aspects of those conditions: 1) the formal relationship between the evaluation and action components, 2) the working relationship regarding research decisions between the evaluation and action components and 3) the working relationship regarding research decisions between the evaluation staff and funding agency staff.

These three variables are particularly important because of the possible implications they have for the quality of the research. Evaluators have spent many hours debating the merits and demerits of conducting evaluations in organizations which are simultaneously administering the action program being evaluated. Our literature is replete with references to the practitioner versus researcher problems[7] and how this threatens objectivity and accessability to data. On the one hand, we hear that evaluators working within the same organization get caught up in the excitement of the program and as such lose their desire and/or ability to do a rigorous scientific evaluation which might then threaten the continuance of that program. The implication is that what emerges is soft testimonial data in place of hard-nosed science. On the other hand, we hear that evaluators operating independent of the action program are viewed by practitioners as 'heartless critics' who intend to build their professional reputations by capitalizing on the program's weaknesses. The implication here is that the practitioner's, acting in self-defense, thwart the researchers by denying them access to the data. Again, the result is a

weak methodological design. While the relationship to the funding agency has not been as popular a topic to debate, there too we see contradictory assertions, some positing it is better to have/more independence, others to have/more interdependence. Rather than making a contribution to that ever increasing 'debate' literature, we turn your attention to our data, first to describe the conditions under which evaluations are actually carried on, and later, in the final section of the paper to ascertain its relationship to quality of research.

Within the total sample population, 38% conducted their evaluation research while working within the same organization that was conducting the action program. Importantly however, fewer than 10% of these 38% were done wherein exactly the same persons administered the action program as those playing a major role in the evaluation research. For example, an experimental program using teaching machines to increase reading was administered by the local community school system. That same school system employed two of its assistant principals who had research training as part of their background to conduct the evaluation study. While the action staff and evaluation staff are composed of different persons, they are common members of a larger umbrella organization. Our data indicated that when the action and evaluation staffs were part of the same organization, they tended to be different rather than the same persons. Of interest here is that when examining this dichotomous variable, which we termed the 'organizational arrangements', we found that of the educationalists, 48% conducted their evaluations within the same organization. Here again, because this is a key variable, we will show this relationship across disciplines. Again we request that you especially note the distribution of psychologists.

Table 2

Organizational Arrangements and Project Director's Academic Discipline

A. Percentaging by Organizational Arrangements

Evaluation and Action by:	Educ.	Psych.	Econ.	Sociol.	Med., S.W and Psycr.	Other	Totals
1. Same Organizations	13.5(12)	34.8(31)	9.0(8)	6.7(6)	20.2(18)	15.7(14)	100.0(89)
2. Different Organizations	9.0(13)	18.1(26)	20.8(30)	15.3(12)	8.3(12)	28.5(41)	100.0(144
Totals	10.7(25)	24.5(57)	16.3(38)	12.0(28)	12.9(30)	23.6(55)	100.0(23

B. Percentaging by Discipline of Project Director

Degree of Project Director	Evaluation and Action Program by Same Organizations	Different Organizations	Totals
Education	48.0(12)	51.0(13)	100.0(25)
Psychology	54.4(31)	45.6(26)	100.0(57)
Economics	21.1(8)	78.9(30)	100.0(38)
Sociology	21.4(6)	78.6(22)	100.0(23)
Med., S.W., & Psycr.	60.0(18)	40.0(12)	100.0(44)
Other	25.5(14)	74.5(41)	100.0(55)
Totals	38.2(89)	61.8(144)	100.0(233)

Note -- $\chi^2_5 = 25.2$, $p < .001$. Three cases omitted because of blank responses.

The second aspect of the structural conditions which we were interested in was the working relationship between the evaluation staff and action staff vis a vis the major research decisions regarding the evaluation. This was especially important insofar as evaluation research literature is replete with references to the implications this variable has for methodological quality. In looking then at 'educationalists' we find that 68% work in relationships

which we categorized as 'joint planning', i.e. where research decisions are made by the two groups in close conjunction with one another. Again, the variable is so important that the table is shown below.

Table 3

Working Relationship between the Action and Evaluation Staffs and the Project Director's Academic Discipline

A. Percentaging by Working Arrangements of Action and Evaluation Staffs

Working Arrangement	Educ.	Psych.	Econ.	Sociol.	Med., S.W. and Psycr.	Other	Totals
Joint Planning	17.2(17)	29.3(29)	7.1(7)	9.1(9)	21.2(21)	16.2(16)	100.0(99)
Action agency reviews	2.2(1)	20.0(9)	20.0(9)	11.1(5)	8.9(4)	37.8(17)	100.0(45)
Eval. agency indep.	7.9(7)	21.3(19)	24.7(23)	15.7(14)	5.6(5)	24.7(22)	100.0(89)
Totals	10.7(25)	24.5(57)	16.3(38)	12.0(28)	12.9(30)	23.6(55)	100.0(233

B. Percentaging by Discipline of Project Director

Degree of Project Director	Joint Planning	Action agency Reviews research	Evaluation Agency indep.	Totals
Education	68.0(17)	4.0(1)	28.0(7)	100.0(25)
Psychology	50.9(29)	15.8(9)	33.3(19)	100.0(57)
Economics	18.4(7)	23.7(9)	57.9(22)	100.0(38)
Sociology	32.1(9)	17.9(5)	50.0(14)	100.0(28)
Med., S.W. & Psycr.	70.0(21)	13.3(4)	16.7(5)	100.0(30)
Other	29.1(16)	30.9(17)	40.0(22)	100.0(55)
Totals	42.5(99)	19.3(45)	38.2(89)	100.0(233)

Note -- $\chi^2_{10} = 36.1$, $p < .001$. Three cases omitted because of blank responses.

Lastly, with respect to the third aspect, i.e. the working relationship between the evaluation staff and funding agency staff vis a vis research decisions, our data indicate that 'educationalists' tend to be more independent (40%) of the funding agency. This table is not presented because the overall relationship between academic discipline of the project director and the working relationship between evaluation and funding was not statistically significant despite the fact that for persons in education there was a discernable pattern.

Before concluding our descriptions of the independent variables examined, one last important dimension was reviewed. In so far as we believed that social action programs should somehow have an overall perspective, theory or frame of reference which guides their program and further, that the presence of a theoretical framework would facilitate an evaluation of that program by helping to define the boundaries or goals, we examined the distribution in responses to this question for the different academic disciplines. First, we should note that of all evaluations done in FY 70 only 18.3% gave evidence of having a formal social structural or social psychological theory guiding their action program. 66% on the other hand stated that their overall perspective was what we called a social service model which in effect translated the notion that 'giving people services was beneficial.' The remaining 16% had no theoretical framework or what was termed an 'a-theoretical naive hypothesis' e.g. senior citizens make good parole officers (with no information to support their assertions). Looking at 'educationalists' and 'theory' then, we find that 16% had a formal structural or social psychological theory, 52% a social service model and 32% no theory at all. In fact, of the total sample, educationalists had the highest proportion in the 'no theory' category Looking at psychologists on the other hand, we find they are very overrepresented in the 'formal theory' group and underrepresented in the 'no theory' group.

In conclusion, to summarize briefly this last section, our data suggest a pattern for 'educationalists' wherein they are very nearly equally divided between conducting evaluations in the same organization as opposed to different organizations with respect to the action component, they lean heavily towards working interdependently with the action staff but independent of the funding agency staff. The pattern is less strong but nonetheless similar for psychologists. With respect to theory 'educationalists' tend to work on action programs without a guiding theory while psychologists are just the opposite.

While the descriptive aspect is most informative, the more interesting and pressing question for us was an analytic one which asked about the quality of the research and how various factors helped to explain some of the variance in that quality. For purposes of our research we defined quality as adherence to the rules of the scientific method, or more specifically the satisfaction of what are considered minimum requirements for good evaluation research.[8] Operationally, we defined quality of research by creating an index composed of several variables which had clustered together when using a factor analytic technique. The alpha for Index A was .42, Index B .53 and Index C 169. Clearly these are not high, but recall that they are three item indices.

Since our data were separated into two distinct areas, i.e. a) the measurement of whether the program was being carried out in accordance with stated specifications and guidelines (input or process measure - Index A) and b) the measurement of impact or change which occurred as a result of the action program (output or effect measure - Index B) we created two separate indices. Additionally because it is asserted[9] that the best evaluative research contains both a measurement of input (process) and a measure of output (effect) we created a third index which was a composite of those variables comprising Indices A and B. This third index is labelled Index C and refers then to what we call 'comprehensive evaluations'. Lastly, it is important to note that

the variables used in these indices were ordered from high quality to low according to the tenets of science and then treated as if they were interval data for purposes of our analyses.[10] A complete explication of our method of index construction as well as the component variables is given in Appendix A.

To compute our results we used a regression analysis format specifically using a particular variation referred to as 'dummy variable regression.'[11] We computed both gross effects and net effects, the former being the total amount of variance in the dependent variable that the independent variable could explain (eta squared or R^2)[12] and the latter being the amount of variance in the dependent variable explained by the independent variable net of (controlling for) all other variables. In computing gross effects we defined each of our indices of quality, Index A, B, and C as our dependent variables, and then regressed each index on each of the independent variables. A test of significance was done for each R^2. Additionally, we computed conditional means, i.e. the mean of the dependent variable for each specific response category of every independent variable. In computing net effects we entered all the independent variables which we had hypothesized would be important into one regression equation and got out an R^2. We then systematically deleted each set of dummy variables for each independent variable, one set at a time, to get the unique variance or net effect of each independent variable. Again, a test of significance was done to determine the statistical significance of each of these net effects. Presented below in Table 4 are the conditional means, gross effects (R^2) and net effects for project director's whose academic discipline was education, as well as for some of the other variables discussed earlier in the paper.

Table 4

Conditional Means and Gross and Net Effects

Variable Name	Conditional Means			Gross Effects			Net Effects		
	Input Quality	Output Quality	Overall Quality	Input Quality	Output Quality	Overall Quality	Input Quality	Output Quality	Overall Quality
Funds Allocated									
1. 10,000 - 49,000	2.29	1.84	4.30	.01	.01	.01	-	-	-
2. 50,000 - 99,000	2.08	1.78	4.02						
3. 100,000 - 149,000	2.19	1.65	3.89						
4. 150,000+	2.20	1.86	3.98						
Nature of Award									
1. Contract	1.95	1.50	3.45	.09***	.09***	.09***	.004	.024*	.027*
2. Grant	2.39	2.04	4.54						
Sponsoring Agency									
1. NIH/NIMH	2.40	2.13	4.65	.06**	.06**	.10**	.005	.016	.005
2. SRS	2.22	1.70	3.43						
3. OE/Other HEW	2.01	1.63	3.67						
4. Justice	2.44	1.81	4.40						
5. HUD/Ag/Labor/OEO	1.96	1.58	3.59						
Focus of Action Program									
1. Health/Mental Health	2.33	1.88	4.35	.03	.03	.05*	.005	.030	.015
2. Education	2.16	1.97	4.18						
3. Income/Security/ Housing/Welfare	2.00	1.58	3.59						
4. Public Safety	2.27	1.84	4.15						
Years of Study									
1. Less than one year	2.48	1.55	4.12	.17***	.13***	.16***	.093***	.044*	.062**
2. 1 to 1 1/2 years	1.72	1.46	3.30						
3. 1 1/2 to 2 years	2.13	1.65	3.82						
4. 2 to 3 years	2.22	1.84	4.05						
5. more than 3 years	2.58	2.28	4.85						
Type of Organization									
1. Profit	1.95	1.52	3.60	.07**	.05*	.06*	.014	.006	.012
2. Non-profit/Research	2.11	1.78	4.03						
3. Educational	2.45	2.02	4.48						
4. Public Service/ Planning	2.15	1.78	3.93						

Table 4 (Continued)

Variable Name	Conditional Means: Input Quality	Conditional Means: Output Quality	Conditional Means: Overall Quality	Gross Effects: Input Quality	Gross Effects: Output Quality	Gross Effects: Overall Quality	Net Effects: Input Quality	Net Effects: Output Quality	Net Effects: Overall Quality
Organizational Arrangements									
1. Eval/Action by Same Organization	2.39	2.08	4.59	.04**	.06***	.09***	.002	.002	.001
2. Eval/Action by Different Organization	2.07	1.62	3.72						
Discipline Project Director									
1. Economics	1.98	1.66	3.71	.08**	.12***	.14***	.027	.044	.062*
2. Psychology	2.55	2.33	4.95						
3. Sociology	2.08	1.43	3.50						
4. Education	2.16	1.65	3.90						
5. Medicine/Psychiatry Soc. Work	2.09	1.72	3.75						
6. Other	2.09	1.64	3.88						
Relationship Between Evaluation and Action									
1. Joint or Action Over Eval.	2.39	1.93	4.43	.06**	.02	.05**	.012	.006	.001
2. Action Reviews Eval. Decisions	1.98	1.77	3.81						
3. Eval. Makes Decisions	2.07	1.66	3.75						
Relationship Between Evaluation and Funding									
1. Funding Makes Eval. Decisions	2.11	1.76	4.11	.01	.03	.03	.013	.012	.025*
2. Joint Decisions	2.03	1.49	3.50						
3. Funding Reviews Eval. Decisions	2.24	1.77	4.12						
4. Eval. Makes Decisions	2.23	1.95	4.21						

Table 4 (Continued)

Variable Name	Conditional Means: Input Quality	Conditional Means: Output Quality	Conditional Means: Overall Quality	Gross Effects: Input Quality	Gross Effects: Output Quality	Gross Effects: Overall Quality	Net Effects: Input Quality	Net Effects: Output Quality	Net Effects: Overall Quality
Theory Guiding Action									
1. Medical or Soc. Service Model	2.13	1.69	3.82	.09***	.08***	.10***	.025*	.005	.011
2. Social Psych/Social Struct. or Cost-Benefit	2.64	2.32	5.02						
3. No Theory	1.90	1.63	3.90						
All Variables Combined				.33***	.32***	.41***			
Grand Means	2.19	1.80	4.08						
Standard Deviations	.75	.89	1.42						

---

rate --

* = .05 level of significance
** = .01 level of significance
*** = .001 level of significance

You will note that net effects have not been computed for budget (funds allocated). This was due to the fact that quite surprisingly the size of the award did not seem to be able to account for any of the variance in the quality of the research (notice the R^2 for funds allocated). Additionally, before constructing our indices we computed measures of association for each of the independent variables e.g. budget, with each of the dependent variables which were later included in each index and there too budget did not seem to vary significantly or systematically with any of the dependent variables. Thus it was not included in the regression equations used to compute net effects.

The indices of quality were constructed such that persons could score from zero to three on Indices A and B and from zero to six on Index C, with higher scores indicating better research quality. Note that on the measurement of input (process) the average quality score was 2.19. While it may appear then that most studies were of fairly high quality it must be emphasized that the index was composed of items which reflected the satisfaction of minimum scientific criteria. For example if in a study assessing the implementation and effectiveness of a remedial reading program, one had taken a simple random sample of both teachers and the target student population, used multivariate statistical techniques to analyze one's data, and thus categorized one's self as having done a quantitative study, a perfect score of 3 would have been attained on Index A. Thus a score of 2.19 on Index A indicates that the quality of evaluation studies in general was not particularly high. In fact only 25% of all evaluations which assessed input (N = 185) scored a 3 while 9% scored a 1 or less. Similarly, for Index B which contained items rating research design, sampling frame and adequacy of measurement, a perfect score of 3 was obtained by only 19% of those who assessed impact (N = 182) even though the mean was 1.80. Additionally, 27% attained a score of 1 or less than 1. The mean for Index C, which was a composite of A and B, and computed

only for those 152 studies which measured process and impact, was 4.08 with 11% attaining a 6 and 22% receiving a score of 3 or less. The median value for Index A is equal to 2.25, for Index B, 2.0, and for Index C, 4.21, which while only slightly higher than the mean scores lends further credence to the assertion that many of the evaluation studies were using less than adequate methodology.

Now, let us examine more closely how well 'educationalists' fare when compared to project directors from other academic disciplines. The eighth entry in Table 4 indicates that psychologists do far better evaluation research than persons from any of the other academic disciplines. On all three of the indices of quality psychologists score roughly one-half standard deviation above the mean. While educationalists turn out to have the second highest scores on two of the three indices, it must be noted that in each of these cases, their scores are slightly below the overall mean, indicating how poorly, relative to psychology, the other disciplines including educationalists are doing. Moreover, to strengthen our findings here, we examined the relationship of the principal investigator's academic discipline to quality of research, as well as that of the person most essential on a daily basis to the evaluation effort and, in both instances found the pattern to be the same. While the order of disciplines other than psychology varied, psychologists remained constant as the single group doing the best evaluation research. And so now we may ask, "Should Evaluation Researchers in Education Have An Inferiority Complex?" The answer, like most answers to scientific questions is unclear. If one's reference group is psychologists, it seems clear that those educationalists doing evaluation research are doing less well. On the other hand, another way of interpreting it is to assert that educationalists doing evaluation research aren't doing it any worse than those in other disciplines, with the exception of psychologists. We suggest that now is the time to recall that some psychologists are educational psychologists and as

such apparently are doing good research. Unfortunately we don't know the number which may be involved here.

One can only speculate as to why psychologists do better research than persons in other disciplines. It is common knowledge that training in psychology for the Ph.D. includes courses in experimental design, statistics and measurement. Additionally, experimentation has a long history in psychology, whereas experimentation is virtually unknown in the field of economics and, sociologists (with the exception of social psychologists) are relatively unfamiliar with it as well. While there has been much emphasis placed on the need for experimentation in the evaluation of educational programs, e.g. Campbell, D. T. and Stanley, J. C. (1963), Trow, M. (1971), etc., our impression is that actual research training in experimentation is still more likely to occur in psychology than in education.

A cautionary note should be interjected here. As the data in Table 4 indicate, only 8% of the variance in Index A, 12% in Index B, and 14% in Index C is explained by the project director's academic discipline. Even more important is to note that most of the effects of this variable can be removed when all the other independent variables are controlled. However we would like to neither understress nor overstress this last statement since we are still in the process of trying to determine causal effects. This is particularly difficult since our own research is obviously not experimental and there is a substantial amount of collinearity among the independent variables. The point is that among our independent variables academic discipline has some of the largest gross effects on quality of research, and yet it explains only a relatively small amount of variation in research quality, and, even these effects can be substantially reduced when other independent variables are controlled.

While time does not permit us here to elaborate the theoretical postulations which might serve to explain why most of the other variables shown in

Table 4 account also for some of the variance in research quality, we can give a very brief summary of what our tentative conclusions are. With respect to what may be termed the 'administrative domain', we note that studies of the highest quality tend to be sponsored by NIH/NIMH, awarded in the form of grants, to evaluate action programs focused on the areas of health or mental health. Regarding what may be termed an 'organizational domain', these same awards tend to be given to persons affiliated with educational institutions who conduct their evaluations within the same organization as the one administering the action program, and who make major research decisions independent of the funding staff and interdependently with the action staff. Lastly, the highest quality seems to be found in studies long in duration, where the evaluations are of action programs guided by some formal theory, and, where the project director's academic discipline is psychology. All of this must be underscored by the fact that there tends to be moderately to high associations between all of these categories of independent variables with one another.

In summation, the general quality of research in evaluation is not high, and the evaluation research done by 'educationalists' identifiable in our sample appears to be not high as well.

Endnotes

1. Suchman, E.A., Evaluative Research, 1967, Russell Sage Foundation, N.Y., pp. 31. See also, Hyman, H., Wright, C., and Hopkins, T., Applications of Methods of Evaluation: Four Studies of the Encampment for Citizenship, 1962, University of California Press, Berkeley, pp. 3.

2. While no discussion appears in this paper, an additional area of investigation taken up was the "Dissemination and Utilization of Evaluation Research". This section was examined and analyzed by Patricia Rieker and will appear in a forthcoming monograph by Freeman, H.E., Bernstein, I.N., and Rieker, P. to be published by the Russell Sage Foundation.

3. Scriven, Michael, "The Methodology of Evaluation", in Perspectives of Curriculum Evaluation, AERA Monograph Series, Rand McNally, 1967, Chicago, pp. 39-83.

4. For a detailed description of the method of sample selection see, Bernstein, Ilene N., Survey of Evaluation Research: A Sociological Examination of a Scientific Activity, unpublished doctoral dissertation, NYU, 1973, chapter on Methodological Procedures.

5. Because of the extreme care we took in selecting our sample we suggest it is unlikely that this figure truly represents persons who merely did not do evaluation. More likely if reflects: a) some persons who were funded for evaluation research but spent the funds for other activities, b) persons who attempt at evaluation but opted to respond not an evaluation rather than answering detailed questions on the methodological procedures of their study, c) an excuse for persons who did not want to take the time to fill out the questionnaire or d) persons who did not understand the directions. Our own hypothesis based on candid conversations with some respondents is that, reason a) above is the principal reason why 26% of those funded in 1970 had not actually done an evaluation study.

6. You will note that this particular table is based on the project director's academic discipline. Suffice it to say that these relationships vary trivially when using the principal investigator's or most essential research person's discipline instead.

7. See for example, Argyris, C., "Creating Effective Research Relationships in Organizations." Human Organization 17:1:34-40., Aronson, S. and Sherwood, C., "Researcher Versus Practitioner." Social Work 12:4:89-96., Coleman, J.S., "Problems of Conceptualization and Measurement in Studying Policy Impacts." Social Science Research Council Conference: Virgin Islands: New York., Evans, J.W., "Evaluating Social Action Programs.", Social Science Quarterly 50:3:568-581., Rodman, H. and Kolodny, R., "Organizational Strains in the Researcher-Practitioner Relationship.", in Francis Caro (ed.), Readings in Evaluation Research. New York: Russell Sage., Ward, D. and Kassebaum, G., "Biting the Hand that Feeds." Paper read at 1966 American Sociological Association Meetings., Weiss, C., "The Politicization of Evaluation Research." Journal of Social Issues. 26:4:57-67., 1971 "Organizational Constraints on Evaluative Research." Bureau of Applied Social Research.

8. Our definition of good evaluation research draws on the work of Suchman, E.A., Evaluative Research, 1967, Russell Sage Foundation, N.Y., Campbell, Donald T., "Reforms as Experiments", American Psychologist, vol. 24, no. 4, April 1969, pp. 409-429, Bernstein, Ilene N. and Sheldon, Eleanor B., "Methods of Evaluative Research", in Social Science Methodology, (ed.) Robert Smith, 1973 Free Press, N.Y., Hyman, Herbert and Wright, Charles, "Evaluating Social Action Programs," in Readings In Evaluation Research, (ed.) Francis Caro, pp. 185-220, 1971, Russell Sage Foundation, N.Y., Freeman, Howard and Sherwood, Clarence, "Research in Large-Scale Intervention Programs", Journal of Social Issues, vol. 21, 1965, pp. 11-28, and Coleman, James S., "Evaluating Educational Programs", The Urban Review, vol. 3, No. 4, February 1969, pp. 6-8, etc.

9. Same studies as those cited above.

10. The positive implications of this method are discussed in Labovitz, Sanford, "Statistical Usage in Sociology: Sacred Cows and Ritual", Sociological Methods and Research, 1972, vol. 1, pp. 13-37.

11. For a discussion of this technique see Cohen, Jacob, "Multiple Regression as a General Data-Analytic System", Psychological Bulletin, 1968, vol. 70, 6, pp. 426-443.

12. Eta squared is equal to the 'between sum of squares' over the 'total sum of squares' as shown in Blalock, Hubert, Social Statistics, 1972, 2nd edition, McGraw Hill, N.Y., pp. 413. Eta squared is also equal to the multiple correlation squared. (R^2)

APPENDIX A

We constructed our indices in the following ways:

Index A - Quality of Measurement of Impact

> Using the variables: Sampling where non-systematic non-random, non-systematic random, and random or non-random cluster samples received a 0, and stratified random, simple random or all observed received a 1, Type of Data Analysis where no statistics, ratings, or impressions received a 0, narratives or impressionistic summaries received a 1, ratings from qualitative data received a 2, simple descriptive statistics received a 3, and multivariate statistics received a 4, and Nature of Data Analysis where qualitative analyses received a 0, an analysis evenly divided between qualitative and quantitative received a 1 and quantitative analyses received a 2,[16]

we weighted each item by $\frac{1}{m-1}$ where m was the number of response categories. For example, an evaluation wherein the measurement of process involved the use of a simple random sample $(\frac{1}{2-1})(1)$, and multivariate statistical procedures $(\frac{1}{5-1})(4)$ and whose analysis was predominantly quantitative $(\frac{1}{3-1})(2)$ would receive a total score of 3 on Index A. That is the total score was equal to the sum of the products of the response code times the weight for each item.

---

[16] While there may be some debate as to the order we have imposed here, i.e. quantitative as higher than half quantitative and half qualitative, we feel justified in so doing since most of the current literature on evaluation research methods, e.g. Suchman, E.A., Evaluative Research, 1967, Russell Sage, N.Y., Caro F. (ed.), Readings in Evaluation Research, 1971 Russell Sage, N.Y., Rossi, P. and Williams W., Evaluating Social Programs, 1972, Seminar Press, N.Y., and Sheldon, E.B. and Bernstein, I.N., "Methods of Evaluative Research", in Social Science Methods, (ed.) Robert Smith, 1973, Free Press, N.Y., strongly suggests that the best evaluations in terms of research quality are those which are highly quantitative.

Index B - Quality of Measurement of Impact or Output:

Using the variables: Nature of Research Design where descriptive studies received a 0, comparative, longitudinal or cross-sectional studies without randomization or control received a 1, experimental designs without both randomization and control received a 2 and, experimental designs with randomization and control groups received a 3 and Representativeness of the Sample where haphazardly drawn samples received a 0, moderately representative samples received a 1, and representative samples received a 2,[17] and Quality of the Measurement Procedures where those responses judged as reflecting adequate measurement received a 1 and less adequate received a 0, an index was constructed. The judgment as to the adequacy of measurement was based on the principles of good measurement [as stated in Attitude Measurement edited by Gene Summers (1969)] and on the apparent good fit between the response given which cited the criteria used on which impact on change were studied and how that was measured. Primarily good fit was assessed on the basis of content validity.[18]

---

[17]This was one of the questions which required a judgmental procedure. The process followed was similar to the coding used on another question concerning the adequacy of sampling. That is, those evaluations which drew simple random or stratified random samples from the populations they wished to generalize their findings to were coded representative or 2. Those which were systematic non-random or random or non-random cluster were coded as moderately representative and those which were non-systematic non-random were coded haphazard. While the first two are clearly defined in Blalock, Hubert M., Social Statistics, N.Y., 1960, McGraw-Hill, pp. 392-410, an example which will help clarify a case of the latter is the following: an evaluation of the effect of a referral program for first offenders was based on a sample of the experience of the first thirty cases referred to the agency.

[18]The satisfaction of a measure having adequate content validity used the definition of content validity as it appears in Kerlinger, Fred, Foundations of Behavioral Research, 1964, N.Y., Holt, Rinehart and Winston, pp. 444-447. An example of a response which was coded adequate was: "the criteria by which the effectiveness of an educational program aimed at increasing cognitive ability of mentally retarded children was the use of standardized reading comprehension, vocabulary, and arithmetic tests, all of which had been pretested for reliability on other similar target populations. Five repeated measures were taken over a 2 year period.

Again, as in Index A, we weighted each item by $\frac{1}{m-1}$ where m was the number of response categories for that item. Thus as in Index A, the total scores for a respondent on Index B ran from zero to three.

Index C - Comprehensive Evaluations: Quality of Measurement of Process and Impact

Using the variables: Sampling, Type of Data Analysis, Nature of Data Analysis, Nature of Research Design, Representativeness of the Sample and Quality of Measurement Procedures, ordered as they were in Indices A and B, we constructed the index by weighting each item by $\frac{1}{m-1}$ where m was the number of response categories for that item. Thus the total scores for a respondent in Index C ranged from zero to six.

BIBLIOGRAPHY

Argyris, C., "Creating Effective Research Relationships in Organizations." Human Organization 17:1:34-40.

Aronson, S. and Sherwood, C., "Researcher Versus Practitioner." Social Work 12:4:89-96.

Bernstein, Ilene N., Survey of Evaluation Research: A Sociological Examination of a Scientific Activity, unpublished doctoral dissertation, N.Y.U., 1973.

Bernstein, Ilene N. and Sheldon, Eleanor B., "Methods of Evaluative Research," in Social Science Methodology (ed.) Robert Smith, 1973, Free Press, N.Y.

Blalock, Hubert, Social Statistics, 1972, 2nd edition, McGraw-Hill, N.Y.

Campbell, D. T., "Reforms as Experiments," American Psychologist, Vol. 24, No. 4, April 1969, pp. 409-429.

Campbell, D. T., and Stanley, J. C., "Experimental and Quasi-Experimental Designs for Research on Teaching," in Handbook of Research on Teaching, (ed.) N. L. Gage, Rand McNally, 1963, Chicago.

Cohen, Jacob, "Multiple Regression as a General Data-Analytic System," Psychological Bulletin, 1968, Vol. 70, 6.

Coleman, J. S., "Problems of Conceptualization and Measurement in Studying Policy Impacts." Social Science Research Council Conference: Virgin Islands: New York.

Coleman, James S., "Evaluating Educational Programs," The Urban Review, Vol. 3, No. 4, February 1969.

Evans, J. S., "Evaluating Social Action Programs." Social Science Quarterly 50:3:568-581.

Freeman, H. E., Bernstein, I. N., and Rieker, P., forthcoming monograph on "Dissemination and Utilization of Evaluation Research," to be published by the Russell Sage Foundation.

Freeman, Howard and Sherwood, Clarence, "Research in Large-Scale Intervention Programs," Journal of Social Issues, Vol. 21, 1965.

Hyman, H., Wright, C., and Hopkins, T., Applications of Methods of Evaluation: Four Studies of the Encampment for Citizenship, 1962, University of California Press, Berkeley.

Hyman, Herbert and Wright, Charles, "Evaluating Social Action Programs," in Readings in Evaluation Research, (ed.) Francis Caro, 1971, Russell Sage Foundation.

Labovitz, Sanford, "Statistical Usage in Sociology: Sacred Cows and Ritual," Sociological Methods and Research, 1972, Vol. 1.

Rodman, H. and Kolodny, R., "Organizational Strains in the Researcher-Practitioner Relationship," in Francis Caro (ed.), Readings in Evaluation Research. New York: Russell Sage.

Scriven, Michael, "The Methodology of Evaluation," in Perspectives of Curriculum Evaluation, AERA Monograph Series, Rand McNally, 1967, Chicago.

Suchman, E. A., Evaluative Research, 1967, Russell Sage Foundation, N.Y.

Trow, Martin, "Methodological Problems in the Evaluation of Innovation" in Readings in Evaluation Research (ed.) Francis Caro, Russell Sage, 1971, N.Y.

Ward, D. and Kassebaum, G., "Biting the Hand that Feeds." Paper read at 1966 American Sociological Association Meetings.

Weiss, C., "The Politicization of Evaluation Research." Journal of Social Issues 26:4:57-67, 1970.

Weiss, C., "Organizational Constraints on Evaluative Research." Bureau of Applied Social Research, 1971.